

# **Hands-On: Running Local LLMs with Docker Model Runner**

**Packt Publication Workshop**

**Rami Krispin, February 15th, 2026**

# Rami Krispin

Senior Manager - Data Science & Engineering

Author | Docker Captain  | LinkedIn Instructor

The AIOps Newsletter

Home Notes GitHub Actions Course Archive About

**ISSUE 7** From Zero to a Dockerized Development Environment in Minutes with GitHub Repository Templates An effective approach for setting up a development environment SEP 21 · RAMI KRISPIN

Running OpenAI GPT OSS Locally with Docker Model Runner and R This is the fourth tutorial in the Docker Model Runner sequence AUG 19 · RAMI KRISPIN

Docker Model Runner - Pull LLMs from Hugging Face This is the third tutorial in the Docker Model Runner sequence AUG 8 · RAMI KRISPIN

Running LLMs Locally with Docker Model Runner and Python This is the second tutorial on the Docker Model Runner sequence JUL 29 · RAMI KRISPIN

**Most Popular**

Getting Started with Docker Running OpenAI GPT OSS Introduction to Docker From Zero to a Dockerized

VIEW ALL



Running LLMs with Docker Desktop  
The coming release of Docker Desktop from Docker, Inc is going t...  
3 min read



The skforecast Project, AI Engineering and New Learnin...  
Happy Saturday! A quick update - the newsletter is moving from...  
3 min read



 The Optuna Project, Advanced Topics in Cryptography, Beyo...  
This week's agenda: Open Source of the Week - the Optuna project New...  
3 min read



# Agenda

- Introduction to DMR
- Setup
- Pull LLMs from Docker Hub & Hugging Face
- Running LLMs via the CLI
- Running LLMs with Python
- OpenCode

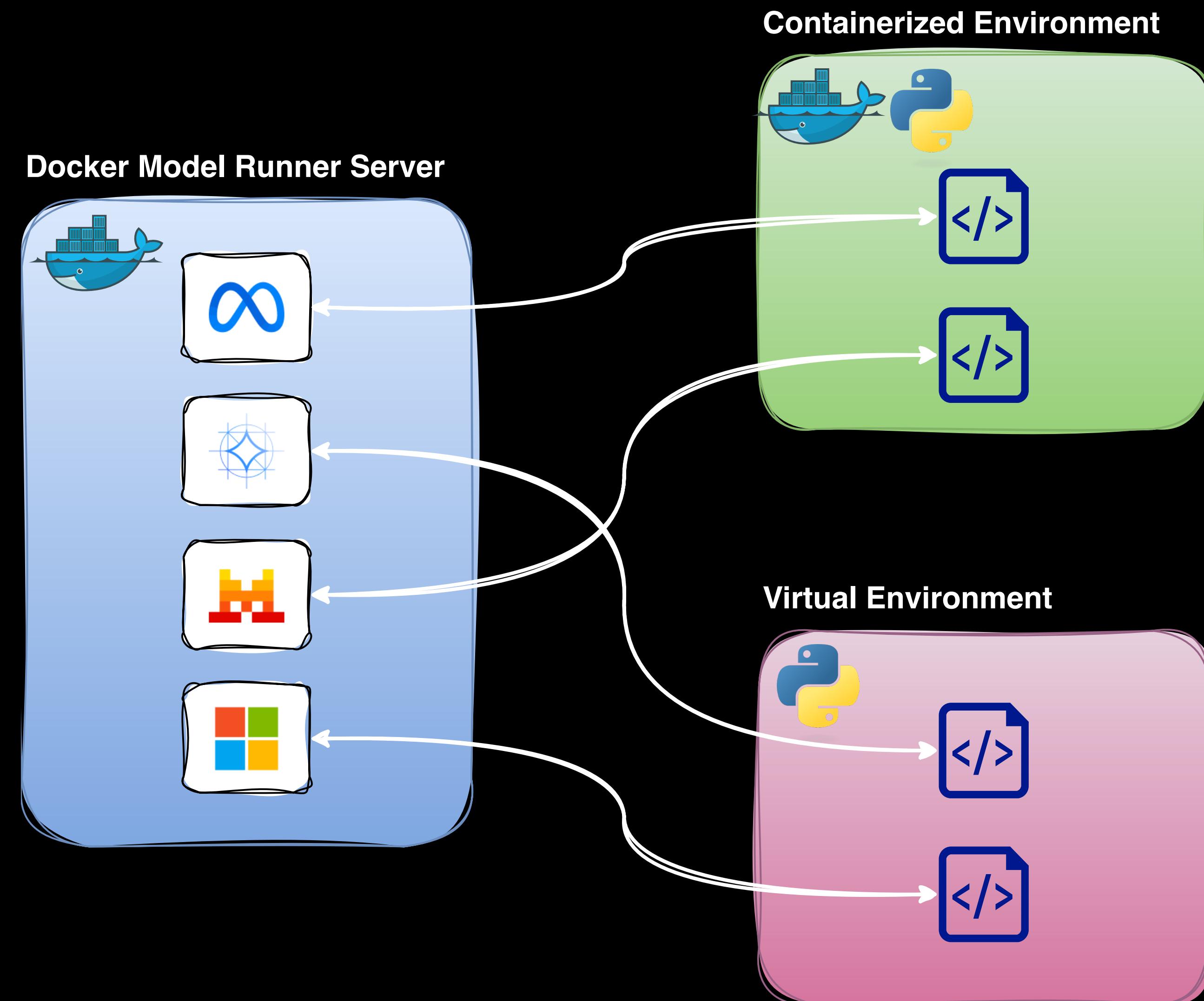
# Poll

- Docker
- Python
- OpenAI Python SDK

[https://github.com/RamiKrispin/  
dmr-workshop-packt](https://github.com/RamiKrispin/dmr-workshop-packt)

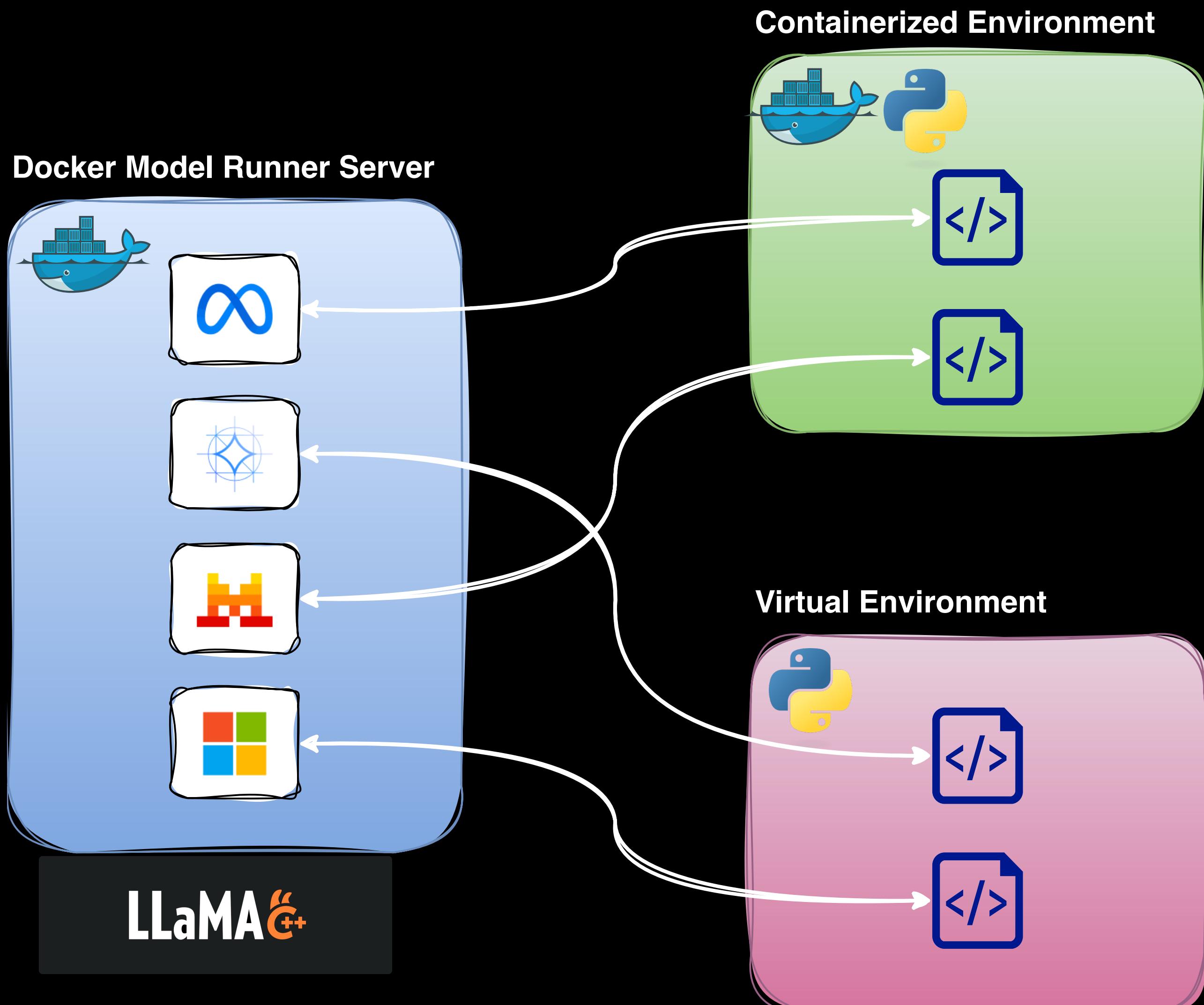
**DMR Doesn't Require Prior Docker Knowledge**

# Introduction to DMR



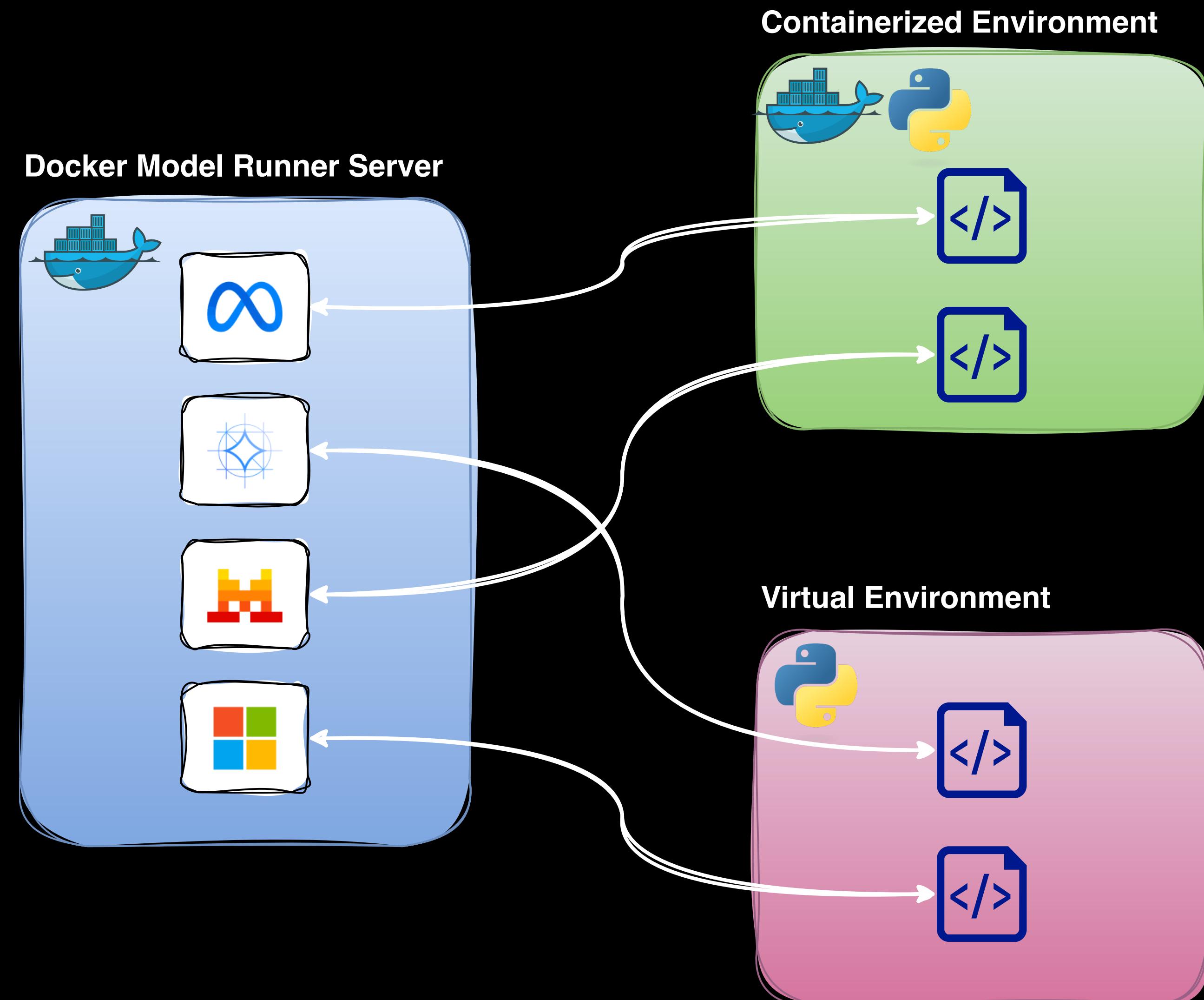
# Introduction to DMR

- Docker Desktop feature
- Based on Llama.cpp
- Easy to setup and use
- Open Container Initiative
- Supports GGUF format
- OpenAI API SDK compatible



# Applications

- Local development of AI applications
- Running local AI applications (local AI agents, code generation, etc.)
- Privacy



# DMR Backend



# General Requirements

- Docker Engine (Linux)
- Docker Desktop 4.40 (Mac) or 4.41 (Windows)
- Docker Hub account
- Hugging Face API key (optional)
- OpenCode/Claude Code (optional)
- Local computational resources

# Setup

# Setup

The screenshot shows the Docker Desktop website and application. At the top, there's a banner with a 'New' badge, a message about Docker + E2B, and links for Docs, Get Support, and Contact Sales. Below the banner is the Docker logo and a navigation bar with AI, Products, Developers, Pricing, Support, Blog, and Company dropdowns, along with Sign In and Get Started buttons. The main heading is 'The #1 containerization software for developers and teams'. A subtext below it says 'Streamline development with Docker Desktop's powerful container tools.' Two buttons at the bottom are 'Choose plan' and 'Download Docker Desktop'. A pink arrow points to the 'Download Docker Desktop' button. At the bottom, a screenshot of the Docker Desktop application interface is shown, featuring a sidebar with 'Containers', 'Images', 'Volumes', 'Builds', and 'Dev Environments' options, and a main panel displaying container usage statistics and a list of running containers.

⚠️ New Docker + E2B. A new partnership bringing trust to AI development. Learn more. → X

Docs Get Support Contact Sales

docker AI Products Developers Pricing Support Blog Company Sign In Get Started

docker desktop

# The #1 containerization software for developers and teams

Streamline development with Docker Desktop's powerful container tools.

Choose plan Download Docker Desktop

docker desktop

Containers Give feedback

View all your running containers and applications. [Learn more](#)

Container CPU usage 0.01% / 1200% (12 CPUs allocated)

Container memory usage 249.6MB / 7.47 GB

Search Only show running containers

Name	Container ID	Image	Port(s)	Last started	CPU %	Actions

# Setup

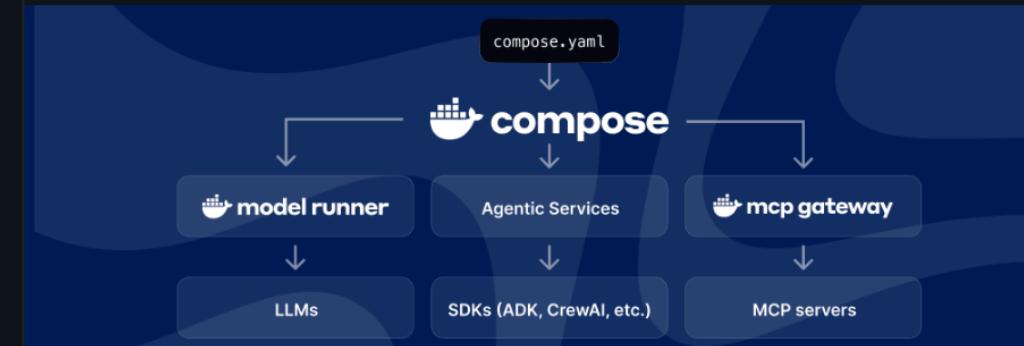


The screenshot shows the Docker Hub homepage. At the top, there's a blue header bar with the Docker Hub logo, a search bar, and navigation links for 'Sign in' and 'Sign up'. Below the header is a large purple banner with the text 'Docker Hardened Images - Secure & Compliant' and a subtext: 'Enterprise-grade Docker images with built-in security, compliance, and continuous updates. Minimize vulnerabilities and deploy with confidence.' A 'Visit catalog now' button is located in the center of the banner.

On the left side, there's a sidebar with several categories:

- Generative AI**
  - AI Models
  - MCP Servers
- Trusted content**
  - Docker Hardened Images
  - Docker Official Images
  - Verified Publisher
  - Sponsored OSS
- Categories**
  - Networking
  - Security
  - Languages & frameworks

The main content area features three spotlight sections:

- AI MEETS COMPOSE**  
**Build AI Agents Faster with Docker Compose**  
Use the workflow you know to develop and deploy across local, cloud, and multi-cloud environments with Docker Compose.  

- MCP**  
**E2B + Docker: Trusted AI**  
Every E2B sandbox includes direct access to Docker's MCP Catalog, a collection of 200+ tools such as GitHub, Perplexity, Browserbase, and ElevenLabs, all enabled by the Docker MCP Gateway.  

- SOFTWARE SUPPLY CHAIN**  
**Secure Your Supply Chain with Docker Hardened Images**  
Use Docker's enterprise-grade base images: secure, stable, and backed by SLAs for Ubuntu, Debian, Java, and more. Regularly scanned and maintained with CVE remediation and long-term support.  


At the bottom, there are sections for 'Machine Learning & AI' and 'Languages & frameworks', each with a 'IMAGE + 1 MODE' button.

# Setup

The screenshot shows the Hugging Face homepage with a dark theme. On the right side, a user profile dropdown menu is open for the user "RamiKrispin". The menu includes options like "Profile", "Notifications", "Inbox (0)", "Trending", "Create organization", "Usage Quota", "Private Storage", "Zero GPU", "Inference Usage", "Get Hugging Face PRO", "Settings", and "Access Tokens". A large yellow arrow points to the "Access Tokens" link. The main content area displays the user's activity feed, which includes posts about updating spaces and liking models.

**Hugging Face** Search models, datasets, users...

Models Datasets Spaces Community Docs Enterprise Pricing

**RamiKrispin**

- + New
- Profile
- Inbox (0)
- Settings
- Billing
- Get PRO

**Organizations**

- Create New

**Resources**

- Getting Started
- Documentation
- Forum
- Tasks
- Learn

Light theme

**My activity**

- All Models Datasets Spaces Papers Collections Community Posts Upvotes Likes Articles

Updated 2 Spaces over 1 year ago

- My Streamlit
- Shiny for Python template

Updated 2 Spaces almost 2 years ago

- Docker Poc
- Shiny for R template

Liked a model about 2 years ago

- Deci/DeciLM-6b-instruct

Trending last 7 d

- deepseek-ai/DeepSeek
- PaddlePaddle/Paddle
- tencent/Hunyuai
- krea/krea-real
- HuggingFaceFW/HuggingFaceFW
- DeepSeek OCR Demo
- veo3.1-fast
- DeepSite v3
- Wan2.2 Animate
- karpathy/fineweb-edu-100b-shuffle
- nick007x/github-code-2025

Profile RamiKrispin

- Notifications
- Inbox (0)

- New Model
- New Dataset
- New Space
- New Collection

Create organization

Usage Quota

Private Storage 0 GB/100 GB

Zero GPU 0/4 min

Inference Usage \$0.00 / \$0.10

Get Hugging Face PRO →

Settings

Access Tokens

Billing

Changelog • 3

Sign Out

# Setup

## opencode

- Intro
- Config
- Providers
- Enterprise
- Troubleshooting

Usage

- TUI
- CLI
- IDE
- Zen
- Share
- GitHub
- GitLab

Configure

- Tools
- Rules
- Agents
- Models
- Themes
- Keybinds
- Commands
- Formatters
- Permissions
- LSP Servers

### Install

The easiest way to install OpenCode is through the install script.

```
curl -fsSL https://opencode.ai/install | bash
```

You can also install it with the following commands:

Using Node.js

NPM    BUN    PNPM    YARN

```
npm install -g opencode-ai
```

Using Homebrew on macOS and Linux

```
brew install sst/tap/opencode
```

Using Paru on Arch Linux

```
paru -S opencode-bin
```

# Setup

```
ramikrispin ~ ⟳ 17:06 ⟳ docker login
Authenticating with existing credentials... [Username: rkrispin]

ℹ Info → To login with a different account, run 'docker logout' followed by 'docker login'

Login Succeeded

ramikrispin ~ ⟳ 17:07 ⟳ hf auth login --token $HF_TOKEN
The token has not been saved to the git credentials helper. Pass `add_to_git_credential=True` in this
function directly or `--add-to-git-credential` if using via `hf` CLI if you want to set the git credential
as well.
Token is valid (permission: fineGrained).
The token `DMR Workshop` has been saved to /Users/ramikrispin/.cache/huggingface/stored_tokens
Your token has been saved to /Users/ramikrispin/.cache/huggingface/token
Login successful.
Note: Environment variable `HF_TOKEN` is set and is the current active token independently from the tok
en you've just configured.
```

# Setup

The screenshot shows the Docker Desktop settings interface in dark mode. The left sidebar lists various sections: General, Resources, Docker Engine, Builders, AI (which is highlighted with a pink rectangle and has a teal arrow pointing to it), Kubernetes, Software updates, Extensions, Beta features, Notifications, and Advanced. The main content area is titled 'AI' and contains the following information:

- Settings for beta AI features can be found [here](#).
- Docker Model Runner**:
  - Enable Docker Model Runner [Give feedback](#)
  - Enable GPU-accelerated inference engines on /var/run/docker.sock and model-runner.docker.internal:80
  - Note: Inference engine support may take a few minutes to initialize.
- Enable host-side TCP support
  - Port: 12434 (highlighted with a yellow box)
  - default: 12434
- CORS Allowed Origins:
  - Custom (dropdown menu)
  - Add origin +

A large blue callout bubble points from the 'Enable host-side TCP support' section towards the right, containing the text: **In case you have NVIDIA GPUs - enable it**.

At the bottom of the window, there are 'Close' and 'Apply' buttons. The status bar at the bottom shows: Engine running, RAM 8.26 GB CPU 0.40%, Disk: 181.52 GB used (limit 1006.85 GB), >\_ ✓ v4.48.0.

# Models

✖ New Docker + E2B. A new partnership bringing trust to AI development. Learn more. → X

hub Search Docker Hub 3K ⓘ ⚡ ⚡ Sign in Sign up

Docker Verified Publisher

Verified Publisher Docker San Francisco, CA, USA https://www.docker.com

**Repositories**

Search by repository name Displaying 1 to 30 of 32 repositories

MODEL

 **ai/granite-docling** Docker

Granite Docling is a multimodal model for efficient document conversion.

Pulls 5.7K Stars 1 Last Updated 17 days

MODEL

 **ai/moondream2** Docker

An open-source visual language model that interprets images via text prompts, fast and powerful.

Pulls 2.9K Stars 0 Last Updated 17 days

MODEL

 **ai/smolvlm** Docker

SmolVLM: lightweight multimodal model for video, image, and text analysis, optimized for devices.

Pulls 4.1K Stars 2 Last Updated 18 days

MODEL

 **ai/granite-4.0-h-small** Docker

32B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.

Pulls 3.1K Stars 1 Last Updated 25 days

MODEL

 **ai/granite-4.0-h-tiny** Docker

7B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.

Pulls 3.1K Stars 2 Last Updated 25 days

MODEL

 **ai/granite-4.0-h-micro** Docker

3B long-context instruct model with RL alignment, IF, tool calling, and enterprise readiness.

Pulls 2.0K Stars 2 Last Updated 25 days

# Models

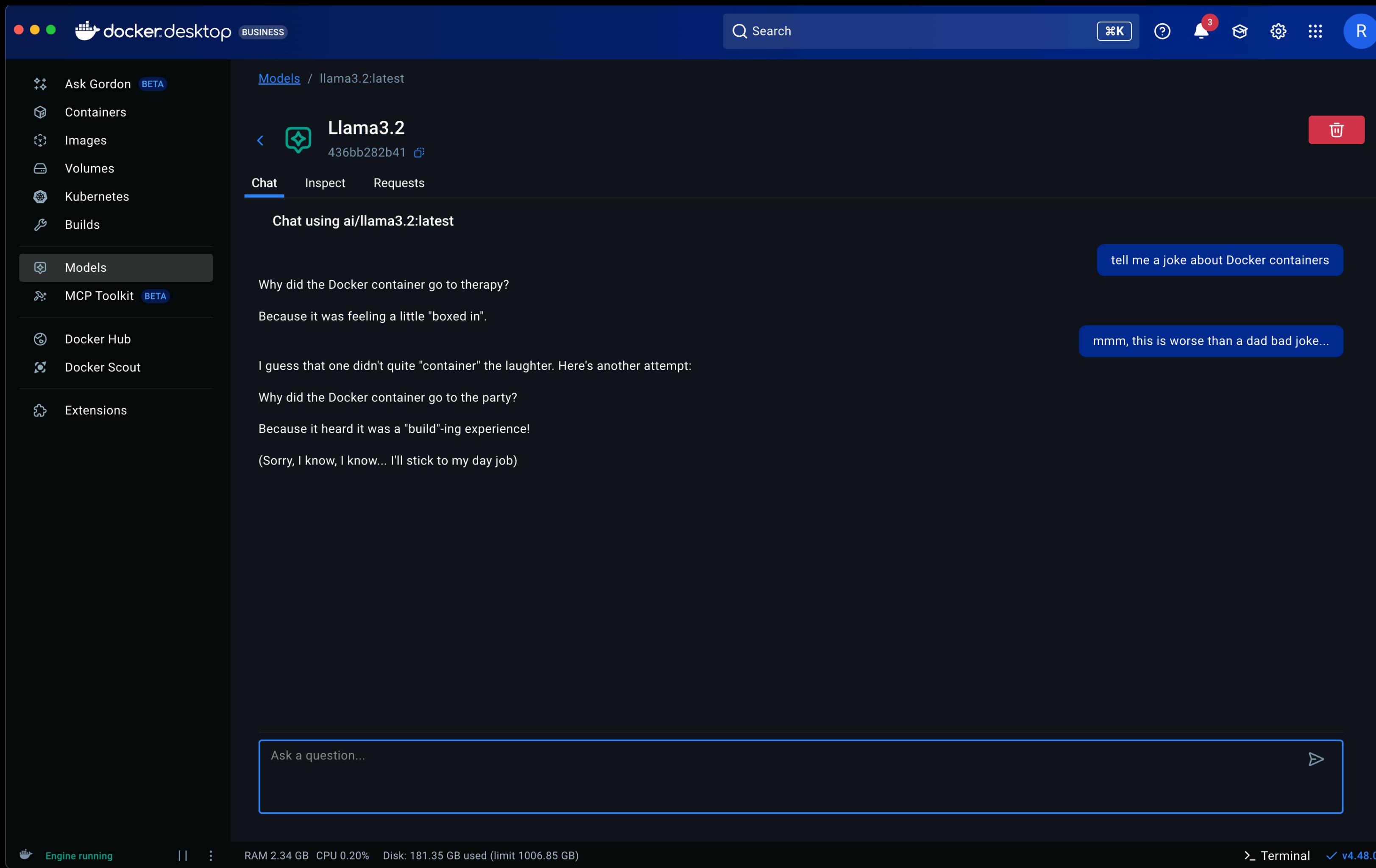
The screenshot shows the Docker Desktop application interface. The left sidebar has a dark theme with white icons and text. The 'Models' option is selected, highlighted with a grey background. Other options in the sidebar include 'Ask Gordon (BETA)', 'Containers', 'Images', 'Volumes', 'Kubernetes', 'Builds', 'MCP Toolkit (BETA)', 'Docker Hub', 'Docker Scout', and 'Extensions'. The main content area is titled 'Models' and features a search bar at the top. Below the search bar, there are tabs for 'Local', 'Requests', 'Logs', and 'Docker Hub', with 'Docker Hub' being the active tab. A sub-header 'Give feedback' is visible above the search bar. The main content displays a grid of nine model cards:

- granite-docling** [Pull](#) (5.7K stars)
- moondream2** [Pull](#) (2.9K stars)
- smolvlm** [Pull](#) (4.1K stars)
- granite-4.0-h-small** [Pull](#) (3.1K stars)
- granite-4.0-h-tiny** [Pull](#) (3.1K stars)
- granite-4.0-h-micro** [Pull](#) (2.0K stars)
- granite-4.0-micro** [Pull](#) (1.4K stars)
- devstral-small** [Pull](#) (3.5K stars)
- magistral-small-3.2** [Pull](#) (7.0K stars)

At the bottom of the main content area, there is a page navigation bar with a central page number '1' and arrows for 'prev', 'next', and 'last'. The footer of the application provides system status: 'Engine running', 'RAM 2.35 GB CPU 0.10%', 'Disk: 181.35 GB used (limit 1006.85 GB)', and a version indicator 'v4.48.0'.

# Hello World!

# Running LLMs via Docker Desktop



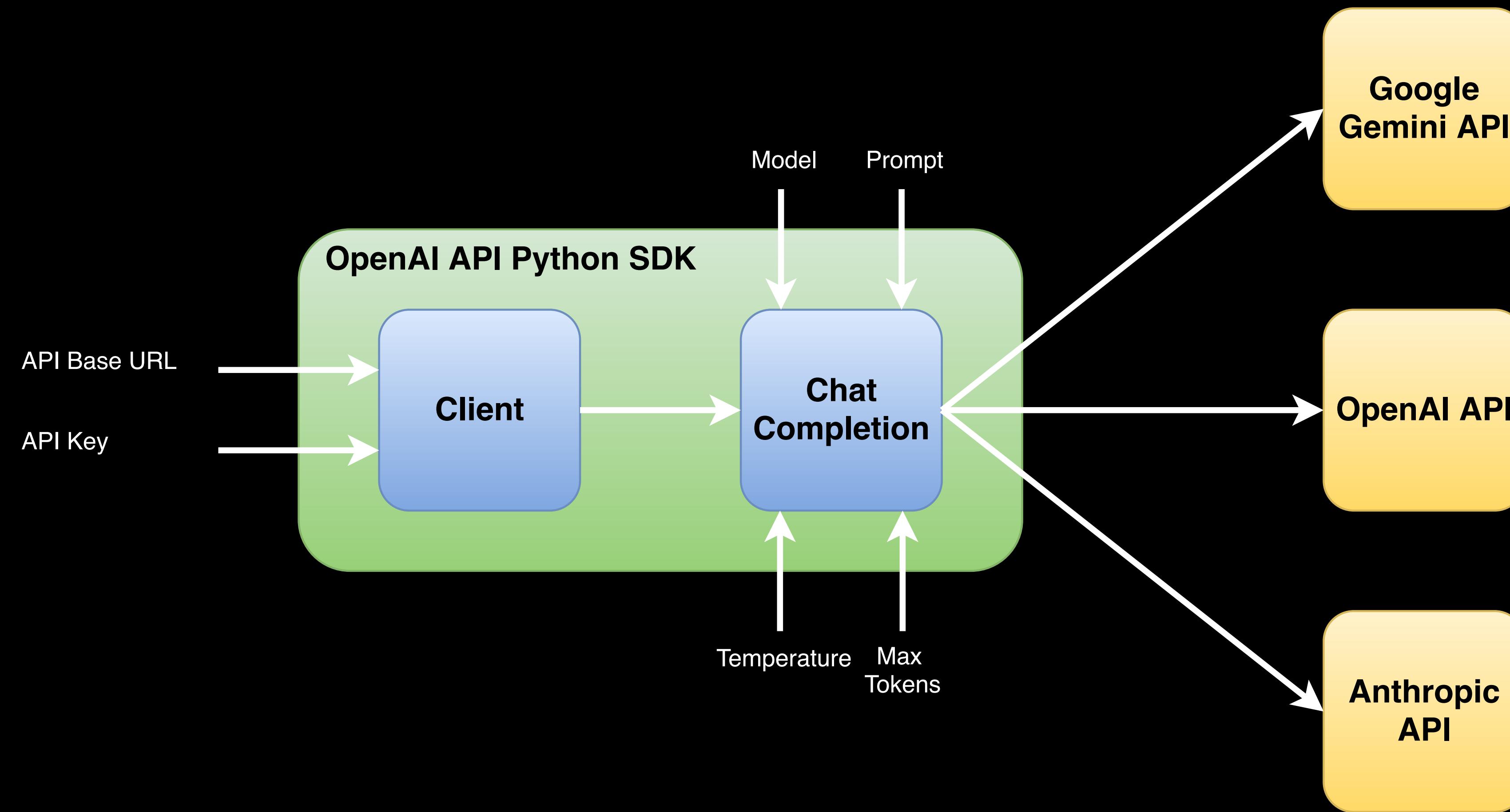
# Running LLMs via the CLI

- CLI core commands
- Running LLMs
- Pull models from Docker Hub
- Pull LLMs from Hugging Face
- Update the context window

# Running LLMs with Python

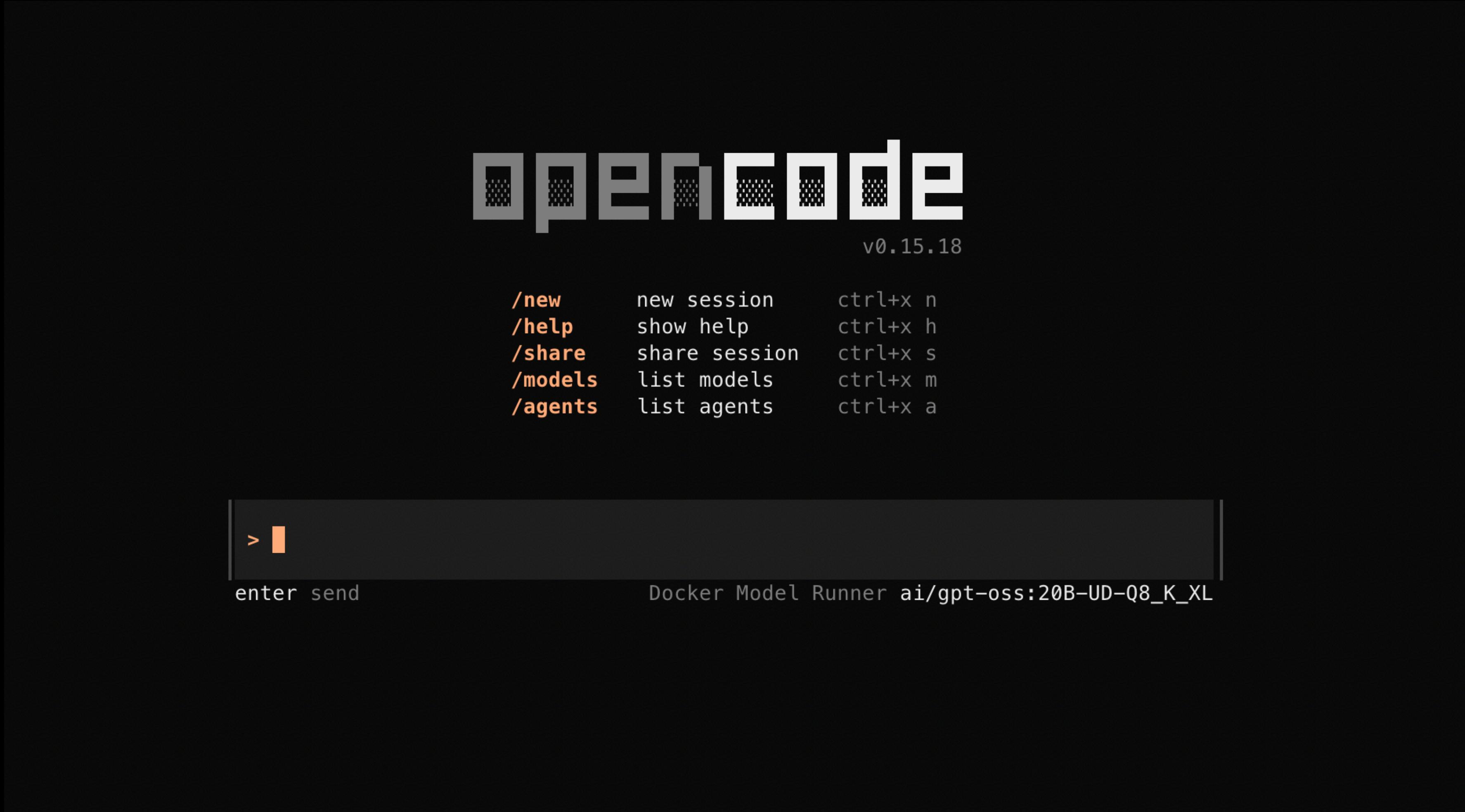
- OpenAI API SDK
- Calling LLMs from Virtual Environment vs. Container
- LangChain

# OpenAI API SDK



# Demo

# OpenCode



# OpenCode

```
{  
  "$schema": "https://opencode.ai/config.json",  
  "provider": {  
    "dmr": {  
      "npm": "@ai-sdk/openai-compatible",  
      "name": "Docker Model Runner",  
      "options": {  
        // Case using inside a container:  
        // "baseURL": "http://model-runner.docker.internal/engines/v1",  
        // Otherwise use the following base URL:  
        "baseURL": "http://localhost:12434/engines/v1",  
        "apiKey": "docker"  
      },  
      "models": {  
        "ai/gpt-oss:20B-UD-Q8_K_XL": {  
          "name": "ai/gpt-oss:20B-UD-Q8_K_XL"  
        },  
        "ai/llama3.2:3B-Q4_0": {  
          "name": "ai/llama3.2:3B-Q4_0"  
        },  
        "ai/devstral-small:24B": {  
          "name": "ai/devstral-small:24B"  
        }  
      }  
    }  
  }  
}
```

# Claude Code

```
Claude Code v2.1.42
Welcome back!
ai/gpt-oss:20B-UD-Q8_K_XL · API Usage Billing
~/Projects/tutorials/odsc-ai-2025-dmr-workshop

Tips for getting started
Run /init to create a CLAUXE.md file with instructions for Claude

Recent activity
No recent activity

/model to try Opus 4.6
) /model

Select model
Switch between Claude models. Applies to this session and future Claude Code sessions. For other/previous model names, specify with --model.

1. Default (recommended)      Use the default model (currently Sonnet 4.5) · $3/$15 per Mtok
2. Opus                      Opus 4.6 · Most capable for complex work · $5/$25 per Mtok
3. Opus (1M context)         Opus 4.6 for long sessions · $10/$37.50 per Mtok
4. Haiku                     Haiku 4.5 · Fastest for quick answers · $1/$5 per Mtok
5. ai/gpt-oss:20B-UD-Q8_K_XL ✓ Custom model

    Effort not supported for ai/gpt-oss:20B-UD-Q8_K_XL

Use /fast to turn on Fast mode (Opus 4.6 only). Now 50% off through Feb 16.

Enter to confirm · Esc to exit
```

# Claude Code

```
export ANTHROPIC_AUTH_TOKEN=docker
```

```
export ANTHROPIC_BASE_URL=http://localhost:12434
```

# Claude Code

claude --model ai/gpt-oss:20B-UD-Q8\_K\_XL

# Resources

- DMR documentation - <https://docs.docker.com/ai/model-runner/>
- Docker podcast - <https://www.youtube.com/watch?v=9XryO1Oxb4A>
- Workshop repo - <https://github.com/RamiKrispin/dmr-workshop-packet>
- The AIOps Newsletter - <https://theaiops.substack.com/>